Gautam Rana

+91-9664513886 gautamelon@gmail.com linkedin.com/in/gautam-rana-rdx github.com/beingPro007

AI Engineer focused on inference optimization, multimodal systems, and model deployment — at the edge and at scale. I ship things that run in production: VLLM pipelines with 15x speedups, CLIP systems that cut 20-second API calls to 30ms, and on-device encoders at 100ms.

Technical Skills

Training & Fine-tuning: PyTorch, HuggingFace Transformers, LoRA/PEFT, CLIP, ViT-14, RealESRGAN
Inference & Deployment: VLLM (PagedAttention), ONNX Runtime, NVIDIA A100 Optimization, Modal Labs
RAG & Search: SPLADE, GCN-based Retrieval, ANNOY Vector Indexing, RAG Pipelines
Infrastructure: Docker, Google Cloud Platform (GCP), Linux, LiveKit
Languages: Python (Advanced), Rust, Go, C++, SQL, Dart
Other: Next.js, Node.js, Flutter, Git/GitHub

Experience

Propelius Technologies — Sep 2024 – Present
AI Engineer — *Surat, Gujarat*

- Implemented a self-learning multilingual product search engine using Amazon's GCN architecture and SPLADE, fine-tuning LoRA adapters on user interaction data to continuously improve retrieval ranking quality.
- Replaced legacy Gemini-based product categorization with a CLIP one-shot pipeline — 400x latency reduction (10–20s → 30ms) and zero per-call API costs, enabling high-volume product ingestion without scaling expenses.
- Deployed Qwen 2.5-VL locally on NVIDIA A100 for document image extraction, eliminating Gemini API dependency and significantly reducing per-document inference cost while preserving output quality.
- Applied VLLM PagedAttention to the document pipeline, reducing per-page inference from 5 minutes to ∼20 seconds (15x speedup) and making high-volume batch processing viable.

Geeks For Geeks — Jul 2024 – Jan 2025
Technical Content Writer — *Remote*

- Authored 20+ technical articles on Git, JavaScript, and System Design, reaching 8,000+ readers.

Projects

Wallee: Edge-AI Asset Discovery | *Flutter, Rust, ONNX Runtime, ViT-14* — Dec 2025 – Present

- Ported ViT-14 patch text encoder to mobile using Rust-based ONNX Runtime, achieving 100–200ms on-device inference with no server roundtrip or cloud cost.
- Designed an optimized asset delivery bridge that reduces package size while enabling full offline functionality.
- Documented the full architecture publicly: How I Built a Lightning-Fast Text Encoder for Edge Devices.

Multimodal AI Agent Platform | *Python, RAG, ANNOY, LiveKit, OpenAI Functions* — May 2025 – Jun 2025

- Built a low-latency voice-to-action platform for real-time multimodal inputs (text, audio, image) routed to domain-specific agents.
- Designed a multi-agent orchestration layer using OpenAI function calling, enabling specialized agents (Legal, Sales) to autonomously handle complex workflows with structured outputs.
- Implemented RAG with ANNOY vector indexing, achieving <300ms retrieval across 5k+ documents.

Education

Uka Tarsadia University — Surat, Gujarat
Bachelor of Science in Computer Science — *Expected: 2026*

Relevant Coursework

- Data Structures
- Machine Learning
- Deep Learning
- RAG Systems
- Computer Networks
- System Design
- Distributed Systems
- Cloud Computing